WebWise: Navigating the Human Genome Project

Kim D. Pruitt

Genome Res. 1997 7: 1038-1039
Access the most recent version at doi:10.1101/gr.7.11.1038

WebWise: Navigating the Human Genome Project

Kim D. Pruitt[1]

National Center for Biotechnology Information (NCBI), National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894 USA

The Human Genome Project has increased the rate of DNA sequence accumulation to the point where information management has become a formidable task. The central repositories for this avalanche of data, GenBank, EMBL (European Molecular Biology Laboratory), and DDBJ (DNA Data Bank of Japan), continue to accumulate DNA sequences at an unprecedented rate. For example, the total number of nucleotides stored in the GenBank database more than doubles every 18 months (Benson et al. 1997). The scientific community is clearly interested in supporting rapid access to high-quality DNA sequence, and, although this remains controversial (Adams and Venter 1996; Bentley 1996), in supporting release of ''unfinished'' DNA sequence data generated by the sequencing centers. (Unfinished DNA sequences generated from a cosmid, BAC, or P1 clone may include nucleotide errors and may consist of unordered or ordered contigs with one or more gaps.) Since the process of ''finishing'' a sequence (which includes resolving any ambiguous bases, contig assembly, gap closure, and annotation) proceeds at a much slower pace than the initial production of sequence, a considerable amount of unfinished sequence can accumulate at the sequencing centers.

Growing interest in timely dissemination of all the data, plus the perception that uneven access to the unfinished DNA sequences could confer an unfair advantage (or disadvantage) to research groups, resulted in increasing pressure on the sequencing centers and international databases to make even unfinished DNA sequence data publicly accessible. This intent was formalized at the International Strategy Meeting on Human Genome Sequencing (held in Bermuda in February 1996) where a consortium of sequencing centers and funding agencies agreed that (1) all publicly funded human sequence data should be promptly released into the public domain, and (2) to promote coordination, sequencing centers should inform a central database of their intent to sequence a given region (''the Bermuda Principles,'' Smith and Carrano 1996). GenBank, EMBL, and DDBJ have supported this agreement by forming a new functional division entirely of unfinished DNA sequence data [high throughput genomic (HTG) sequence] (Ouellette and Boguski 1997); to support the coordination effort, these international databases are developing web sites to collect and display information concerning the chromosomal regions targeted for sequencing. [The Human Genome Sequence Index (HGSI) will be available from NCBI's home page: http://www.ncbi.nlm.nih.gov/.] On an individual level, however, the sequencing centers themselves have risen to the task of making their data fully available by establishing and maintaining World Wide Web sites. In fact, these web sites have become an integral component of the ongoing sequencing effort. A sequencing center's web site is often the best place to find an integrated overview of the mapping and sequencing progress, the DNA sequence, and future plans for a particular region of interest. The growing importance of these web sites to the sequencing community was made apparent at the recent Hilton Head meeting when several speakers referred to their center's web site as a further source of information (Ninth International Genome Sequencing and Analysis Conference. September 13–16, 1997. Hyatt Regency, Hilton Head, SC).

Not surprisingly, the rapid increase in DNA sequence data has been matched by a proliferation of web sites to disseminate, discuss, or ''link to'' the actual DNA sequence information. This plethora of web sites, and the wealth of information available there, is a powerful research tool—but one that is likely underused. It can be incredibly time-consuming and confusing to fully utilize this resource even for an experienced web navigator, as it is often difficult to maneuver through the maze of sites to find the relevant information. Indeed, if you do not have the URL address at hand, it can even be difficult to locate a particular web site. The web has become so large that it is often challenging to phrase a useful search query to locate a particular site. For example, a recent search engine query for ''Human Genome Project'' *and* ''sequencing center'' yielded 43 matches, of which only 2 were to a sequencing center listed in Table 1. The initial difficulty in identifying the correct web site to look at is compounded by the fact that the different sequencing centers have employed a variety of organizational strategies for their web sites. While variety is the hallmark of the web, the lack of any organizational standards can make it even more time-consuming to find data, as it is frequently not at all apparent how to navigate around a web site. Nor is it always obvious exactly what resources are available at a given site. This WebWise series of articles, of which this is the first, is meant to be a navigational aid for sequence sites available on the web.

The WebWise series will review the Human Genome Project sequencing centers' web sites. In addition to simply pointing the way to the different centers, this series will provide an outline of each center's organizational strategy, discuss the type of information available there, and evaluate the general ease of use. There are many web sites that pro-

[1]E-MAIL pruitt@myrtle.nlm.nih.gov; FAX (301) 435-2433.

7:1038–1039 ©1997 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/97 $5.00

Table 1. Human Genome Project DNA Sequencing Center Web Sites

Sequencing center	URL
· Baylor College of Medicine Human Genome Sequencing Center	http://gc.bcm.tmc.edu:8088/home.html
· Lawrence Berkeley National Laboratory Human Genome Center	http://www-hgc.lbl.gov/
· Lawrence Livermore National Laboratory	http://www-bio.llnl.gov/bbrp/genome/genome.html
· Stanford Human Genome Center	http://www-shgc.stanford.edu/
· The Genome Sequencing Center Jena	http://genome.imb-jena.de/
· The Institute for Genomic Research (TIGR)	http://www.tigr.org/
· The Sanger Centre	http://www.sanger.ac.uk/
· U. of Oklahoma Advanced Center for Genome Technology	http://www.genome.ou.edu/
· U. of Texas SW Medical Center Genome Center	http://mcdermott.swmed.edu/datapage/
· U. of Washington Genome Center	http://www.mbt.washington.edu/
· Washington U. School of Medicine Genome Sequencing Center	http://genome.wustl.edu/gsc/gschmpg.html
· Washington U. Center for Genetics in Medicine	http://genome.wustl.edu/cgm/cgm.html
· Whitehead Institute/MIT Center for Genome Research	http://www-genome.wi.mit.edu/

vide either human genome mapping data or a limited amount of sequence data; however, the scope of the series is to review the larger sequencing centers, listed in Table 1, which are the prime contributors to the Human Genome Project. Although many of these web sites include information pertaining to several different organisms, this review series will primarily emphasize the portion of each web site relevant to the Human Genome Project, but the overview will give a good sense of the other information available at each site.

One of the main objectives of this series is to provide a handy resource to the research community. To achieve this goal, each article will have a uniform organization to facilitate rapid assessment of a given web site's features in relation to previously discussed sites. Each WebWise review will focus on a couple of the sequencing centers and will include a general overview of each site, general features, and any special features. To further enhance the utility of this series, the content of each site will be presented in a tabular synopsis using a simple rating system that will allow rapid comparison of web sites. Each tabular summary should enable the reader to quickly ascertain the format and accessibility of sequence data, software availability, search capability, update frequency, and general ease of use for a given sequencing center's web site.

The sequencing center web sites are one of the most powerful tools currently available for accessing sequence data, assessing progress, and gaining early information related to ongoing work within every genetics laboratory. Making these sites more easily accessible to individuals who are less familiar with the inner workings of these sites will serve to increase the knowledge in the community at a rate that will hopefully equal the amount of data currently flowing into these sites.

REFERENCES

Adams, M.D. and J.C. Venter. 1996. *Science* 274: 534–536.

Benson, D.A., M.S. Boguski, D.J. Lipman, and J. Ostell. 1997. *Nucleic Acids Res.* 25: 1–6.

Bentley, D.R. 1996. *Science* 274: 533–534.

Ouellette, B.F.F. and M.S. Boguski. 1997. *Genome Res.* 7: 952–955.

Smith, D., and A. Carrano. 1996. *Hum. Genome News* 7: 19.